FORECASTING SURVIVAL BY SAMPLE REUSE TECHNIQUES

by

Seymour Geisser
University of Minnesota

Technical Report No. 333

December, 1978

FORECASTING SURVIVAL BY SAMPLE REUSE TECHNIQUES

by

SEYMOUR GEISSER[1]

University of Minnesota

0. Introduction.

Methods relevant to forecasting in survival or reliability data situations are generated by the predictive sample reuse technique in partial conjunction with a Bayesian approach. Initially we assume the entire fine structure of an exponential survival distribution cum gamma prior distribution on the exponential parameter. Subsequently the predictive distribution of a future observation from the process is obtained. In the gamma prior we essentially assume one of the hyperparameters known (or guessed) and the other unknown. An estimate for the latter is produced by the predictive sample reuse method. The question of censored data, where ambiguity exists in the execution of the predictive sample reuse method, is tentatively resolved by the ploy of pseudo-observations that are supplied from a Bayesian or other structure.

The utilization of the approximate predictive distribution, i.e., with one hyperparameter estimated, is valid to the extent of the appropriateness of the fine structure assumptions with uncertainty commensurate with the roughness of the approximation. On the other hand the predictor itself may be useful considerably beyond the bounds of the initial structure assumed in that it may be robust as a point predictor for a variety of possible structures. Further it may be most useful in a low structure situation, where any specific distributional assumptions are fraught with peril.

[1]Work supported by NIH grant 1R01 GM25271

1. Predictive Sample Reuse.

The predictive sample reuse method was presented in a variety of detailed forms, Geisser (1974, 1975a), Stone (1974). Here we shall delineate it in a very simple manner appropriate to the particular applications that flow from it under discussion in later sections.

Suppose we have a set of observations $x^{(N)} = (x_1, \ldots, x_N)$ and we are interested in predicting a future observation from the process generating observations of this kind. We further assume a predictive function used to forecast a potentially observable value

$$(1.1) \qquad x_{N+1} = f(x^{(N)}, \alpha); \ \alpha \in \Omega$$

where α is defined as some unknown constant or set of such unknowns whose domain is Ω. Next we define an average discrepancy function

$$(1.2) \qquad D(\alpha) = N^{-1} \sum_{j=1}^{N} a_j(\alpha) d(x_j, f_j)$$

where $d(x_j, f_j)$ represents a discrepancy between the observed value x_j and the predictive function f_j which is formed as in (1.1) except that x_j has been excised from f and $a_j(\alpha)$ represents a weight for the j^{th} descrepancy that may depend on α. In other words each observation x_j has been withheld in turn and the predictive function formed excluding it is used to forecast x_j. Then $D(\alpha)$ is minimized for values of α restricted to Ω which we assume yields a unique value $\hat{\alpha}$. This leads to the predictor

$$\hat{x}_{N+1} = f(x^{(n)}, \hat{\alpha}) = \hat{f}.$$

For a more detailed exposition of the method involving multiple observational omissions and various schemata of omission, as well as applications, see Geisser (1974, 1975a, 1976).

In applying this method to survival or reliability data, it is quickly apparent that an inherent deficiency exists. The method as stated depends on the full knowledge of the sample values. But for this type of problem quite often our knowledge for a portion of the sample is restricted by the fact that the observations were censored at particular values. In order to remedy this lack of knowledge of fully observed values we introduce pseudo-observations. They depend on α and are determined in this instance from conditional predictive functions. In less structured situations other means may be necessary to establish reasonable values for pseudo-observations. Two procedures utilizing a pseudo-observation approach are presented. The first proposal substitutes the pseudo-observations into the discrepancy measure prior to minimization. This leads rather naturally to considering schemes whereby the censored observations are weighted differently than uncensored ones as opposed to previous applications where $a_j(\alpha) = 1$ and the data were inherently fungible. Of course there could arise situations where a sample of uncensored observations may require different weights because of a decision as to their treatment or a model for their generation. Here, even though we start with a scheme that treats the observations exchangeably the approach of fitting the censored observations into the predictive sample reuse framework naturally induces consideration of differential weighting schemes.

A second proposal involves the substitution of the pseudo-observations into the solutions as if all the values were fully observed and solving the requisite algorithm. This approach is discussed in detail in section 4.

2. Exponential Survival.

Suppose we have a random sample $X_1,\ldots,X_N$ on an exponential random variable X whose density is

$$(2.1) \qquad f(x|\mu) = \mu e^{-\mu x} \quad \mu > 0, \quad x > 0 .$$

If our prior objective or subjective information is subsumed in a prior density for μ,

$$(2.2) \qquad p(\mu) \propto \mu^{\delta-1} e^{-\gamma\mu}, \quad \gamma > 0, \quad \delta > 0$$

and we are interested in predicting a value x_{N+1} for the random future observation X_{N+1} given the previous N observations $x^{(N)}$, say, then the predictive density for X_{N+1} is easily calculated to be , for $x_{N+1} > 0$,

$$(2.3) \qquad f(x_{N+1}|x^{(N)}) = \int p(\mu|x^{(N)})\, f(x_{n+1}|\mu)^{N+1}\, d\mu$$

$$= (N + \delta)(N\bar{x} + \gamma)^{N+\delta}/(N\bar{x} + \gamma + x_{N+1})^{N+\delta+1}$$

where $\bar{x}$ is the sample mean and $p(\mu|x^{(N)})$ is the posterior density of μ given the previous N observations $x^{(N)}$. Hence our forecast about X_{N+1} involves the hyperparameters γ and δ which enter the problem via the distribution of the parameter μ. Before any observations are taken one can also find the predictive (marginal) density of the generic variable X, namely

$$(2.4) \qquad f(x) = \int f(x|\mu)p(\mu)d\mu$$

$$= \delta\gamma^{\delta}/(\gamma + x)^{\delta+1}, \quad x > 0 .$$

Hence it is convenient and perhaps more appropriate to think about these hyperparameters in terms of predicting X before any observations are

taken rather than in how they modulate the assumed prior distribution of μ. Therefore, prior to the sample, we have

$$
\begin{aligned}
&E(X) = \gamma/(\delta - 1) = g \\
&\mathrm{Var}(X) = \delta\gamma^2/(\delta - 2)(\delta - 1)^2 = g^2(1 + \alpha)/(1 - \alpha)
\end{aligned}
\tag{2.5}
$$

where $\alpha = (\delta - 1)^{-1}$.

Clearly $\mathrm{Var}(X)$ exists for $0 < \alpha < 1$, and $E(X)$ exists for $\alpha > 0$ while the distribution exists for all $\alpha \notin [-1,0]$. Hence if one could frame his prior opinions about the potentially observable values of X in terms of its expectation and variance then one can easily execute the whole predictive process by solving for the appropriate values δ and γ from (2.5) and substituting them in (2.3).

It is to be noted that (2.3) and (2.4) were obtained from (2.1) and (2.2). However, for the predictivist who would prefer to start from (2.1) and (2.4) in terms of convenience of framing his predictions this is somewhat awkward. Interestingly enough in this case starting with $f(x|\mu)$ and $f(x)$ is sufficient to obtain $p(\mu)$ and $f(x_{N+1}|\bar{x})$, which is a more logical and appealing approach for the predictivist. This is true here because $f(x)$ is the unique Laplace transform of $\mu^{-1}p(\mu)$.

Now as we mentioned previously making all of these assumptions yields the requisite information for making probability statements about a future value provided that one has specified values for g and α. However while one may often be willing to hazard a guess at g, one may be far less willing to specify a value for α.

We now shall apply the predictive sample reuse method in order that the data itself should yield a value for α once g has been assumed.

If we had already observed $X^{(N)} = x^{(N)}$ and wished to predict a future value for X_{N+1}, we could use the posterior expectation of X_{N+1} obtained from the predictive density given by (2.3). This is easily calculated to be

$$(2.6) \qquad E(X_{N+1}) = (N\bar{x} + \gamma)/(N + \delta - 1) = (\alpha N\bar{x} + g)/(\alpha N + 1) = f \ .$$

Note that when $\delta \to 1$ and $\gamma \to 0$, we obtain the usual predictor $\bar{x}$.

In terms of the predictive sample reuse method, Geisser [1975], equation (2.6) may be utilized as a predictive function. In order to supply a value for α we apply the method using one-at-a-time omissions and a squared discrepancy as follows: The average squared discrepancy is

$$(2.7) \qquad D(\alpha) = N^{-1} \sum_i (f_i - x_i)^2 = N^{-1} \sum_i \left[\frac{\alpha(N-1)\bar{x}_i + g}{\alpha(N-1) + 1} - x_i \right]^2$$

where f_i and $\bar{x}_i$ are defined respectively as the predictive function and the sample average with x_i omitted. In order to find a suitable α we minimize $D(\alpha)$ with respect to α for $\alpha \geq 0$. (Note again that for the density given by (2.4) $Var(X)$ exists only for $0 < \alpha < 1$ although the distribution for X exists for $\delta > 0$ and hence for all $\alpha \notin [-1,0]$. Nevertheless we shall restrict ourselves to $\alpha > 0$ since this is essentially the range on α for which the prior mean exists.)

We can easily evaluate

$$(2.8) \qquad D(\alpha) = [(N - 1)s^2(\alpha N + 1)^2 + N(g - \bar{x})^2]/N[\alpha(N - 1) + 1]^2 \ ,$$

where $s^2 = (N - 1)\sum_{i=1}^{N}(x_i - \bar{x})^2$. Taking the derivative with respect to α and setting this equal to zero yields the solution

$$(2.9)\qquad \begin{aligned} &\hat{\alpha} = (t^2 - 1)/N && \text{for } t^2 > 1 \\ &\hat{\alpha} \to 0 && \text{if } t^2 \le 1 \end{aligned}$$

where $t^2 = N(g - \bar{x})^2/s^2$. Hence this yields the predictor

$$(2.10)\qquad \begin{aligned} &f(\hat{\alpha}) = \hat{f} = [(t^2 - 1)\bar{x} + g]/t^2 && \text{for } t^2 > 1 \\ &f(\hat{\alpha}) \to g && \text{if } t^2 \le 1 . \end{aligned}$$

Of course for the strict Bayesian the use of $\hat{\alpha}$ and its derived value $\hat{\delta}$ contradicts the fundamental canon of Bayesianism that the prior hyperparameters should not depend on the data. However it should serve as an approximate solution to the problem in the sense that the unknown hyperparameter δ is replaced by $\hat{\delta}$ if $\hat{\alpha} > 0$ in (2.3), given the high structure assumptions. This problem and method for solution was first proposed by Geisser (1975b) with further commentary, Geisser (1976, 1977).

It may also be mentioned that the predictor $\hat{f}$ can also be conceived as totally independent of the Bayesian process and the likelihood when obtained from this approach in the sense that we have merely chosen f as a point predictor for X_{N+1} and have ascertained $\hat{f}$ by a squared discrepancy measure. We also note that the predictive function f is basically a linear combination of the mean $\bar{x}$ and the prior guess g with weights αN and 1. There are other Bayesian models which can lead to forecasting the next observation as linear combinations of a prior mean and the sample mean when the predictive expectation of a future observation is utilized. In this regard then one could define a predictive function that is a linear combination of the mean and a guessed value g,

(2.11) $\quad f^* = \alpha^* \bar{x} + (1 - \alpha^*)g \qquad 0 \le \alpha^* \le 1 \ .$

This yields, for squared discrepancy and one-at-a-time omissions, Geisser (1975a),

$$\hat{\alpha}^* = (t^2 - 1)/[t^2 + (N-1)^{-1}] \quad \text{for} \quad t^2 > 1 \ , \qquad (2.12)$$
$$= 0, \qquad \text{if} \quad t^2 < 1 \ .$$

Hence

$$\hat{f}^* = [(t^2 - 1)\bar{x} + N(N-1)^{-1}g]/[t^2 + (N-1)^{-1}], \quad \text{for} \quad t^2 \ge 1 \qquad (2.13)$$
$$= g \qquad \text{if} \quad t^2 < 1 \ .$$

Clearly $\alpha^* = \alpha N/(\alpha N + 1)$ if $t^2 < 1$ in terms of the transformed predictive function. On the other hand $\hat{\alpha}^* < \hat{\alpha}N/(\hat{\alpha}N + 1)$, for $t^2 > 1$, the estimation procedure not being invariant under such a transformation. However they will be quite close as they are asymptotically equivalent for large N. Comparison of $\hat{f}$ with $\hat{f}^*$ reveals they are also converge for large N, but slightly more weight is attached to $\bar{x}$ in $\hat{f}$ than in $\hat{f}^*$.

In summary then, in the assumed presence of the high initial structure, f should be preferable but for robustness to other structures leading approximately to the aforementioned linear combination, f^* may be preferable. In any event the difference is negligible for large N. In the absence of any distributional assumptions both predictors are viable methods for having something to say about the prediction of future observations.

3. Censored Data.

In many cases of survival or reliability studies the experiment is usually terminated before all of the subjects or units have expired or failed. Suppose the experiment is such that for d of the observations, failure times are recorded as $x_1,\ldots,x_d$, while the remaining $N - d$ observations have survived but were censored at values $x_{d+1},\ldots,x_N$. Hence

$$L(\mu) = \prod_{i=1}^{d} f(x_i|\mu) \prod_{i=d+1}^{N} [1 - F(x_i|\mu)]$$

where $F(x_i|\mu)$ is the distribution function of X_i. For the exponential case, clearly

$$(3.1) \qquad L(\mu) \propto \mu^d \, e^{-\mu[d\bar{x}_d+(N-d)\bar{x}_{N-d}]}$$

where $\bar{x}_d = d^{-1} \sum_1^d x_i$ and $\bar{x}_{N-d} = (N-d)^{-1} \sum_{i=1}^{N-d} x_{d+i}$. From (3.1) and (2.2) we can obtain first the posterior density of μ and then as previously the predictive density for a future observation X_{N+1},

$$(3.2) \qquad f(x_{N+1}|x^{(d)}, x^{(N-d)})$$

$$= (d+\delta)(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma)^{d+\delta}/(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma + x_{N+1})^{d+\delta+1}$$

where $x^{(d)}$ represents the observations whose failure times are recorded and $x^{(N-d)}$ the censored observations. Further the predictive expectation, to be used as the predictive function, is

$$(3.3) \qquad E(X_{N+1}) = [d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma]/(d + \delta - 1)$$

$$= [(d\bar{x}_d + (N-d)\bar{x}_{N-d})\alpha + g]/(\alpha d + 1) = f \, .$$

Note that for $\delta \to 1$ and $\gamma \to 0$ we obtain the usual predictor $\bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d}$. Due to censoring there is difficulty in appropriately executing the predictive sample reuse method. One tentative solution is to generate $N - d$ pseudo-observations having values x'_{d+i}, $i = 1,\ldots,N-d$, say. These are the presumed failure times for the censored observations $x_{d+1},\ldots,x_N$. We shall take as the pseudo value x'_{d+i}, the expectation of the predictive distribution of X_{d+i} given $X_{d+i} > x_{d+i}$, the censored value. More precisely the likelihood in (3.1) is used but with x_{d+i} omitted, i.e., based on all the observations but x_{d+i}. This is then combined with the prior density of μ whence the posterior density of μ is obtained and subsequently the predictive density of X_{d+i} computed. From this we then compute the conditional density of X_{d+i} given $X_{d+i} > x_{d+i}$,

$$(3.4) \qquad f(x|X_{d+i} > x_{d+i}) = \frac{(d+\delta)(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma)^{d+\delta}}{(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma + x - x_{d+i})^{d+\delta+1}} .$$

Further computation yields

$$(3.5) \qquad E(X_{d+i}|X_{d+i} > x_{d+i}) = [(d+\delta-1)x_{d+i} + d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma]/(d + \delta - 1)$$

$$= x_{d+i} + \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d})\alpha + g}{\alpha d + 1} = x'_{d+i} ,$$

and

$$(3.6) \qquad Var(X_{d+i}|X_{d+i} > x_{d+i}) = \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma)^2(d+\delta)}{(d+\delta-1)^2(d+\delta-2)} = \frac{d + \delta}{d+\delta-2} f^2$$

the latter being independent of i.

Now in executing the sample reuse method with predictive function given by(3.3)using the actual observations $x_1,\ldots,x_d$ and the pseudo-observations $x'_{d+1},\ldots,x'_N$ given by (3.5) it seems sensible to give the pseudo-observations a weight that differs from that assigned to the uncensored observations in contradistinction to an unweighted and consequently inadequate solution, Geisser (1975b). We note that

$$(3.7) \qquad \mathrm{Var}(X_i|\mu) = \mu^{-2} \qquad \text{for} \quad i = 1,\ldots,d \ .$$

Since μ is unknown we shall compute

$$(3.8) \qquad E_\mu[\mathrm{Var}(X_i|\mu)] = E_\mu[\mu^{-2}]$$

over the posterior distribution of μ. This results in

$$(3.9) \qquad E_\mu(\mu^{-2}) = \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma)^2}{(d+\delta-1)(d+\delta-2)} = \frac{d+\delta-1}{d+\delta-2}\ f^2$$

where f is as defined in (3.3).

We can define a weighted discrepancy for $d > 1$, $N-d > 1$ as follows:

$$(3.10) \qquad D(\alpha) = E_\mu^{-1}(\mu^{-2}) \sum_{j=1}^{d} \left[\frac{[(d-1)\bar{x}_{d,j} + (N-d)\bar{x}_{N-d}]\alpha + g}{\alpha(d-1)+1} - x_j\right]^2$$

$$+ [\mathrm{Var}(X|X > x_{d+i})]^{-1} \sum_{k=d+1}^{N} \left[\frac{[d\bar{x}_d + (N-1-d)\bar{x}_{N-d,k}]\alpha + g}{\alpha d + 1} - x'_k\right]^2 ,$$

where $\bar{x}_{d,j}$ and $\bar{x}_{N-d,k}$ are respectively the sample means of d-1 uncensored observations omitting x_j and the mean of N-1-d censored observations omitting x_k.

After some algebraic manipulation we obtain

$$(3.11)\qquad D(\alpha) = \frac{(d-1)s_d^2(\alpha d+1)^3 + d(g-\bar{x}_d + \alpha(N-d)\bar{x}_{N-d})^2(\alpha d+1)}{[\alpha(d-1)+1][(d\bar{x}_d+(N-d)\bar{x}_{N-d})\alpha+g]^2}$$

$$+ \frac{[\alpha(d+1)+1][\alpha(d-1)+1]}{[(d\bar{x}_d+(N-d)\bar{x}_{N-d})\alpha+g]^2} \cdot \sum_{j=d+1}^{N} x_j^2$$

where $(d-1)s_d^2 = \sum_{j=1}^{d}(x_j - \bar{x}_d)^2$.

The solution then for α is obtained by differentiating (3.11) with respect to α and setting it equal to zero. This will result in a polynomial in α, whose roots are stationary points. After discarding negative and complex roots, the positive roots α, say, need be compared with $D(0)$ and $D(\infty)$ to ascertain the global minimum for $\alpha \geq 0$.

For $d = 1$ and $N > 2$ only the second term in (3.11) obtains and formal minimization in this case yields $\alpha = \infty$, so that $\hat{f} = N\bar{x}$, the usual predictor in this case.

For $d > 1$ and $N = d + 1$ only the first term in (3.11) obtains. Minimization then follows in the same manner as in the discussion for $d > 1$ and $N-d > 1$.

It is to be noted that in the weighting we merely used terms that reflected variation. Perhaps a more appropriate weighting scheme would also include covariation among those values that are correlated. As a step in this direction we can take cognizance of the covariance among the pseudo-observations.

A simple calculation reveals that the joint predictive density of X_{d+i} and X_{d+j} $i \neq j = 1, \ldots, N-d$ conditional on $X_{d+i} > x_{d+i}$ and $X_{d+j} > x_{d+j}$ is

$$(3.12) \quad f(z,w|X_{d+i} > x_{d+i},\ X_{d+j} > x_{d+j}) = \frac{(d+\delta)(d+\delta+1)(d\bar{x}_d+(N-d)\bar{x}_{N-d}+\gamma)^{d+\delta}}{(d\bar{x}_d+(N-d)\bar{x}_{N-d}+z-x_{d+i}+w-x_{d+j})^{d+\delta+2}}$$

whence we calculate

$$(3.13) \quad \mathrm{Cov}(X_{d+i}, X_{d+j}|X_{d+i} > x_{d+i},\ X_{d+j} > x_{d+j}) = (d+\delta)^{-1}\ \mathrm{Var}(X_{d+i}|X_{d+i} > x_{d+i}) \quad \text{for } i \neq j,\ i,j = 1,\ldots,N-d.$$

Use of this alters the second term in (3.11) to

$$(3.14) \quad \frac{[\alpha(d+1)+1][\alpha(d-1)+1]}{(\alpha N+1)(\alpha d+1)[(d\bar{x}_d+(N-d)\bar{x}_{N-d})\alpha+g]^2}\left[\alpha(N-1)+1)\sum_{j=1}^{N-D} x_{d+j}^2 - 2\alpha\sum_{i>j}^{N-D} x_i x_j\right]$$

When as is often the case all of the N-d observations are censored at the same value, say x_o, then (3.14) simplifies

$$(3.15) \quad \frac{[\alpha(d+1)+1]^2[\alpha(d-1)+1]\ (N-d)x_o^2}{(\alpha N+1)\ [(d\bar{x}_d+(N-d)x_o)\alpha + g]^2}\ .$$

This term is then $[\alpha(d+1)+1]/[\alpha N+1]$ times the second term in (3.11), indicating roughly the diminished effect of the contribution of the portion of $D(\alpha)$ involving the pseudo-observations by taking into account their covariance structure. Of course this further complicates arriving at a solution for α and it is not clear just how significant the resulting improvement would be.

The most complex weighting scheme would also attempt to take into account covariation between uncensored observations and pseudo-observations. Now for $i=1,\ldots,d$, $j=d+1,\ldots,N$; $X_j' = X_j + \bar{X}_d + (N-d)d^{-1}\bar{X}_{N-d}$

$$(3.16)\quad \mathrm{Cov}(X_i, X_j' \mid \mu) = \frac{\alpha}{\alpha d+1}, \quad V(X_i \mid \mu) = \frac{\alpha\mu^{-2}}{\alpha d+1} .$$

Again using (3.9) we find that

$$(3.17)\quad E_\mu[\mathrm{Cov}(X_i, X_j' \mid \mu] = \frac{f^2}{d+\delta-2}$$

Hence we may use as a weighting matrix the inverse of the N x N partitioned matrix

$$(3.18)\quad V = \frac{f^2}{d+\delta-2} \begin{pmatrix} (d+\delta-1)I & J_{12} \\ J_{21} & (d+\delta-1)I + J_{22} \end{pmatrix} \begin{matrix} d \\ N-d \end{matrix}$$

where J_{ij} is a matrix all of whose entries are unity. The inverse is readily computed as

$$(3.19)\quad V^{-1} = \frac{[\alpha(d-1)+1]}{(\alpha d+1)f^2} \begin{pmatrix} I + \dfrac{\alpha^2(N-d)}{(\alpha d+1)^2 + \alpha(N-d)} J_{11} & \dfrac{-\alpha^2}{(\alpha d+1)^2 + \alpha(N-d)} J_{12} \\ \dfrac{-\alpha^2}{(\alpha d+1)^2 + \alpha(N-d)} J_{21} & I - \dfrac{\alpha}{(\alpha d+1)^2 + \alpha(N-d)} J_{22} \end{pmatrix}$$

Now for $d > 1$ and $N-d > 1$, let

$$(3.20)\quad \Delta_j = f_j - x_j \quad \text{for } j = 1,\ldots,d$$

$$= f_j - x_j' \quad \text{for } j = d+1,\ldots,N$$

where again f_j is the predictive expectation f omitting the j^{th} observation. Further, letting $\Delta' = (\Delta_1,\ldots,\Delta_N)$ we can now define

$$D(\alpha) = \Delta' V^{-1} \Delta$$

and minimize this for $\alpha > 0$. Again evaluation of $D(\alpha)$ leads to rather complicated algebra which we shall omit.

Once a solution $\hat{\alpha}$ is rendered we can convert it to obtain the approximate predictive distribution of a future observation or just use $\hat{f}$ as a point predictor.

For the second kind of predictive function

$$(3.21) \quad f^* = \alpha^*(\bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d}) + (1-\alpha^*)g = \alpha^* h + (1-\alpha^*)g$$

which does not lean as much on the previous high structure assumptions, we use as pseudo-observations

$$(3.22) \quad x'_{d+i} = x_{d+i} + \bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d} = x_{d+i} + h$$

This is akin to frequentist prediction since using x'_{d+i}, $i = 1,\ldots,N-d$ as actual observations in conjunction with $x_1,\ldots,x_d$ preserves the frequentist predictor, $\bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d}$, as this is the average of both uncensored values and pseudo-observations. Now (3.22) can also be obtained by letting $\delta \to 1$ and $\gamma \to 0$ in (3.5).

Here the simplest weighted squared discrepancy measure neglecting covariation but not variances is

$$(3.23) \quad D(\alpha^*) \propto \sum_{j=1}^{d}(f_j^* - x_j)^2 + \frac{d}{d+1}\sum_{j=d+1}^{N}(f_j^* - x'_j)^2$$

where f_j^* is f^* as in (3.21) but with x_j omitted. The weighting here is again closer to a frequentist approach although it also can be obtained from (3.6) and (3.9) by letting $\delta \to 1$. Let $f_j^* = \alpha^* h_j + (1-\alpha^*)g$ so that

$$(3.24) \qquad \begin{aligned} h_j &= (d-1)^{-1}(d\bar{x}_d + (N-d)\bar{x}_{N-d} - x_j) && \text{for } j = 1,\ldots,d \\ &= \bar{x}_d + d^{-1}[(N-d)\bar{x}_{N-d} - x_j] && \text{for } j = d+1,\ldots,N \end{aligned}$$

then the minimization of $D(\alpha^*)$ with respect to α^* yields

$$(3.25) \qquad \begin{cases} \hat{\alpha}^* = \dfrac{\sum_{j=1}^{d}(h_j-g)(x_j-g) + d(d+1)^{-1}\sum_{j=d+1}^{N}(h_j-g)(x_j-g)}{\sum_{j=1}^{d}(h_j-g)^2 + d(d+1)^{-1}\sum_{j=d+1}^{N}(h_j-g)^2} & \text{for } 0 \le \hat{\alpha}^* \le 1 \\ = 1 & \text{for } \hat{\alpha}^* > 1 \\ = 0 & \text{for } \hat{\alpha}^* < 0. \end{cases}$$

If one uses a scheme with no weighting at all then

$$(3.26) \qquad \begin{cases} \hat{\alpha}^* = \dfrac{N(h-g)^2+(h-g)(d-1)^{-1}d^{-1}(N-d)\bar{x}_{N-d}-(d-1)^{-1}d^{-1}(N-d)^2x^{-2}_{N-d}-s_d^2-d^{-1}\sum_{d+1}^{N}x_j^2}{(d+1)(h-g)^2+2(d-1)^{-1}(h-g)d^{-1}(N-d)\bar{x}_{N-d}+(d-1)^{-2}d^{-1}(N-d)^2x^{-2}_{N-d}+(d-1)^{-1}s_d^2+d^{-2}\sum_{d+1}^{N}x_j^2} \\ = 0 \qquad \text{if } \hat{\alpha}^* \le 0 \\ = 1 \qquad \text{if } \hat{\alpha} \ge 1 \end{cases}$$

A slightly different solution can be obtained by altering the function h. Previously h was defined as the sum of all the observations censored and uncensored, devided by the number of uncensored observations. We also noted that h was the mean of the uncensored values and the pseudo-observations.

Hence we could change the definition of h to this mean value which keeps invariant the value of the predictive function for given α. However h_j would now be altered to

$$(3.27)\quad \begin{cases} h_j' = (N-1)^{-1}[N\bar{x}_d + \frac{(N-d)N}{d}\bar{x}_{N-d} - x_j] & \text{for } j = 1,\ldots,d \\ = \bar{x}_d + (N-d)d^{-1}\bar{x}_{N-d} - (N-1)^{-1}x_j & \text{for } j = d+1,\ldots,N\ . \end{cases}$$

The solution for α^* is now obtained by substituting h_j' for h_j in (3.25).

An unweighted solution in this case is, Geisser 1975b,

$$(3.28)\quad \begin{cases} \hat{\alpha}^* = \dfrac{N(g-h)^2 - A}{N(g-h)^2 + (N-1)^{-1}A} & \text{for } \hat{\alpha} > 0 \\ = 0 & \text{for } \hat{\alpha} \le 0 \end{cases}$$

where

$$(3.29)\quad (N-1)A = (d-1)s_d^2 + d^{-1}(N-d)^2\bar{x}_{N-d}^2 + \sum_{j=d+1}^{N} x_j^2\ .$$

In both (3.24) and (3.27) it is required that $d > 1$ and $N-d > 1$. If $d = 1$ and $N > 2$ then the solution for α^* is the ratio of the second terms in (3.25) utilizing either h_j or h_j' respectively. For $d > 1$, $N = d+1$, the solution is the ratio of the first terms.

4. An Alternative Algorithm

Another approach to censored data sets using sample reuse techniques will now be described in somewhat more general terms than necessarily indicated by the problem at hand. Let $X = (x_1,\ldots,x_d)$ and $X^* = (x_{d+1},\ldots,x_N)$ represent respectively the completely and partially observed data sets--with the understanding that the observable x_j for $j > d$ represents incomplete information of some kind on a realization of the random variable X_j. Let $Y = (y_{d+1},\ldots,y_N)$ represent the set of values which were partially observed as X^*, i.e. we suppose for the moment that we had actually fully observed X^* so that the values would be Y, say. We then compute a complete solution for α, say $\tilde{\alpha} = \tilde{\alpha}(X,Y;Z)$ in the usual fashion e.g. in the previous example this would be (2.9). But we need values for y_j the components of Y. We now assume a condtional predictive function for the components of Y,

$$y_j = \hat{x}_j(X,X^*,Z,\alpha) = x'_j(\alpha)$$

e.g. in the previous case this would be (3.5). Now let $X^*(\alpha)$ represent the set of values inserted for Y, i.e. for each component y_j we insert $x'_j(\alpha)$. Lastly we then have the algorithm

$$\alpha = \tilde{\alpha}(X, X^*(\alpha),Z)$$

which needs be solved for α. Call the solution $\hat{\alpha}$ and one can use this in the previous work either to predict a future observation conditionally or unconditionally.

We now apply this to the censored situation of the previous section. Using (2.9)

$$\tilde{\alpha} = \frac{t^2(\alpha) - 1}{N} \tag{4.1}$$

where from (3.5)

$$(4.2) \qquad x_j'(\alpha) = x_j + \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d})\alpha + g}{\alpha\, d+1} \qquad j > d.$$

Let

$$(4.3) \qquad \bar{x}(\alpha) = \frac{1}{N}\left[\sum_{j=1}^{d} x_j + \sum_{j=d+1}^{N} x_j'(\alpha)\right] = \bar{x} + \frac{(N-d)}{N}\left(\frac{N\bar{x}\alpha + g}{\alpha d + 1}\right)$$

where $N\bar{x} = \sum_{\alpha=1}^{N} x_j$. Let

$$(4.4) \qquad \beta = \frac{N-d}{N}\left(\frac{N\bar{x}\alpha + g}{\alpha d+1}\right)$$

$$(4.5) \qquad (N-1)s^2(\alpha) = \sum_{j=1}^{d}(x_j-\bar{x}-\beta)^2 + \sum_{j=d+1}^{N}(x_j + \beta - \bar{x} - \beta)^2$$

$$= (N-1)s^2 + d\beta^2 - 2\beta d(\bar{x}_d - \bar{x})$$

where $(N-1)s^2 = \sum_{j=1}^{N}(x_j-\bar{x})^2$. Now by definition

$$(4.6) \qquad t^2(\alpha) = \frac{N(\bar{x}(\alpha) - g)^2}{s^2(\alpha)} .$$

Hence substituting (4.6) in (4.1) and solving for α in terms of β i.e.,

$$(4.7) \qquad N\alpha + 1 = \frac{(N-d)(g-\bar{x}-\beta)}{d\beta - (N-d)\bar{x}}$$

we obtain a quadratic equation in β

$$(4.8) \qquad a\beta^2 + b\beta + c = 0$$

where

$$(4.9) \qquad \begin{cases} a = \dfrac{d(N^2-d)}{N-1} \\ b = 2(N-d)d(\bar{x}-\bar{x}_d)(N-1)^{-1} + dN(\bar{x}-g) - N(N-d)\bar{x} \\ c = (N-d)s^2 + N(N-d)\bar{x}(g-\bar{x}) . \end{cases}$$

After obtaining the solution $\hat{\beta}$ we solve for $\hat{\alpha}$ from (4.7).

This approach has the advantage of simplicity--both in terms of treating observations fungibly, as it were, and yielding simpler solutions.

References

Geisser, S. (1971). The inferential use of predictive distributions. Foundations of Statistical Inference (B.P. Godambe and D.A. Sprott, eds.) 456-69. Toronto, Montreal: Holt, Rinehard and Winston.

Geisser, S. (1974). A predictive approach to the random effect model. Biometrika, 61, 101-107.

Geisser, S. (1975a). The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70, 350, 320-328.

Geisser, S. (1975b). Bayesianism, predictive sample reuse, pseudo-observations, and survival. Bulletin of the International Statistical Institute 40, 3, 285-289.

Geisser, S. (1975c). A new approach to the fundamental problem of applied statistics. Sankhya, 37, B, 4, 385-397.

Geisser, S. (1976). Predictivism and sample reuse. Proceedings of the 21st Design of Experiments Conference, pp. 385-397.

Geisser, S. (1978). A predictivistic primer. Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys. (in press)

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. (with Discussion). J. Roy. Statist. Soc. B. 36, 111-147.